PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
G06F 11/00, 11/20

A1

(11) International Publication Number: **WO 99/03038**

(43) International Publication Date: 21 January 1999 (21.01.99)

(21) International Application Number: PCT/US98/14451

(22) International Filing Date: 13 July 1998 (13.07.98)

(30) Priority Data:
08/891,539 11 July 1997 (11.07.97) US

(71) Applicant: MARATHON TECHNOLOGIES CORPORATION [US/US]; 1300 Massachusetts Avenue, Boxboro, MA 01719 (US).

(72) Inventors: LORD, Christopher, C.; 294 Hubbardston Road, Princeton, MA 01541 (US). SCHWARTZ, David, B.; 88 Flanagan Drive, Framingham, MA 01701 (US).

(74) Agent: HAYDEN, John, F.; Fish & Richardson P.C., 601 Thirteenth Street, N.W., Washington, DC 20005 (US).

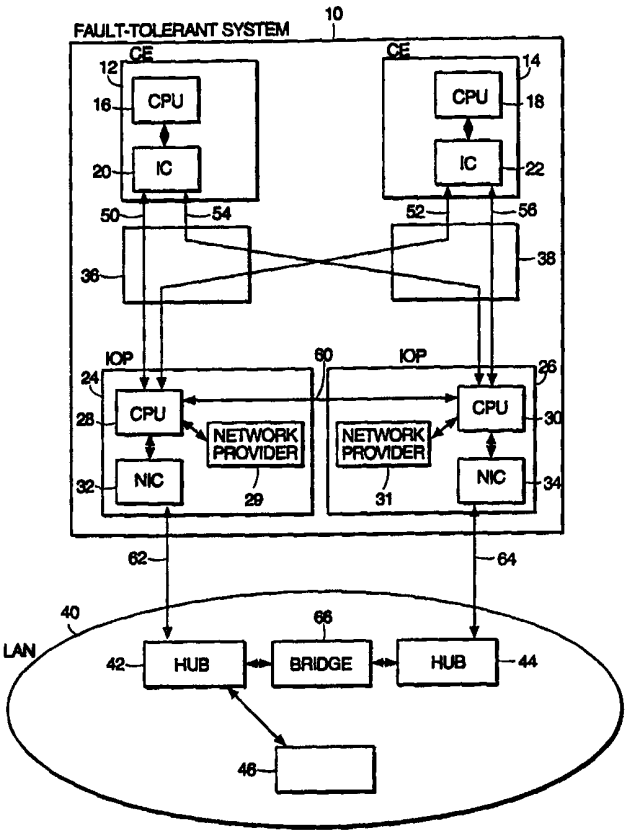
(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: ACTIVE FAILURE DETECTION

(57) Abstract

Failures in a fault–tolerant computer system which includes two or more input/output processors connected to a data communication system are detected by monitoring data communication. The computer system is able to detect failures associated with a primary input/output processor, as well as with a standby input/output processors, and is also able to discriminate between failures of the input/output processors and communication failures in the data communication network itself. In addition to using hearbeat–like transmissions, various other categories of data communication are also used to detect failures. The system is able to detect failures when the input/output processors are on a common network segment, allowing the processors to monitor identical data traffic, as well as when the processors are on different segments where, as a result of filtering behavior of netwok elements such as active hubs, the processors may not be able to monitor identical data traffic.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

ACTIVE FAILURE DETECTION

Background

This invention relates to the detection of failures, such as communication failures, in a fault-tolerant computing system.

Redundant hardware elements are commonly used in fault-tolerant computing systems. Individual elements of the system typically attempt to detect faults by monitoring signals generated by other elements in the system or generated externally to the system.

In addition, an element of the system may periodically transmit a so-called "heartbeat" signal that indicates proper operation of the element. If the heartbeat signal is not received by another element in the system, the receiving element can suspect that the transmitting element is not operational. However, failure to receive a heartbeat signal also may result from a fault in the communication path between the two elements. In general, fault handling should distinguish between a fault in an element of the system and a fault in the communication path between elements.

Redundant network interface controllers (NICs) are used in fault-tolerant computing systems to provide reliable, uninterrupted communication with an external network. In general, one NIC operates in a primary, or active, mode in which the NIC is responsible for communication with other devices on the network, while the other NIC operates in a standby mode.

In operation, the NICs can exchange heartbeat messages to detect failures in a path from one NIC through the external network and back to another NIC. A failure in the path between NICs can occur at several points, including the input or output stages of the NICs, the transmitting or receiving connections between the NICs and the external network, or in the external network

itself. The point of connection to the external network is generally at a port of a network hub, with the hub being connected to multiple network devices. Each NIC may be connected to a different hub in the external network to avoid having a single hub become a critical point of failure.

Summary

The invention provides detection of failures in a fault-tolerant computer system that includes two input/output processors connected to a data communication system. The computer system is able to detect failures associated with a primary input/output processor, as well as with a standby input/output processors. The system also is able to discriminate between failures of the input/output processors and communication failures in the data communication network itself. The system analyzes categories of data communication other than "heartbeat-like" transmissions to detect failures. The system is able to detect failures when the input/output processors are on a common network segment that allows the processors to monitor identical data traffic. The system also is able to detect failures when the processors are on different segments where, as a result of filtering behavior of network elements such as switches or active hubs, the processors may be unable to monitor identical data traffic.

A timing criterion may be applied to a category of data communications processed by each input/output processor and a relationship between results obtained for each processor may be used to detect a failure. For example, a failure may be indicated when a difference between the timing of data communication exceeds a threshold. The timing criterion can be the time of last transmission or reception of a category of communication.

A category of data communication can be, for example, messages originating outside the fault-tolerant system, such as from another computer system coupled to the data communication system. These messages may be addressed to a group of systems of which the fault-tolerant system is a member. The category also may include messages originating from one of the input/output processors, including messages addressed to the other processor, or messages originating from some other element of the system.

In one aspect, generally, the invention features detecting a failure in a fault-tolerant computer system that includes a first input/output processor and a second input/output processor coupled to a data communication system. A timing criterion is applied to a category of data communications processed by the first and second input/output processors to produce first and second timing results. A relationship between the timing results is determined, and whether a failure has occurred is detected based on the determined relationship.

Embodiments of the invention may include one or more of the following features. For example, detecting whether the failure has occurred may include determining that a failure has occurred when a difference between the timing results exceeds a threshold value.

The timing criterion may be a time of last transmission or reception. The category of data communications may include messages originating from the first input/output processor, such as messages directed to an address to which the second input/output processor is normally responsive or messages sent from the first input/output processor and directed through the data communication system to the second input/output processor. The category of data communications also may include messages originating outside the computer system,

such as messages originating at a second computer coupled to the data communication system, or messages addressed to a group of systems of which the computer system is a member. The category of data communications also may include messages originating from a third element of the computer system in data communication with the input/output processors.

The timing criterion may be applied to the category of data communications processed by the first input/output processor at the first input/output processor, and the first timing result may be sent from the first input/output processor to the second input/output processor. The timing criterion may be applied to the category of data communications processed by the second input/output processor at the second input/output processor, and the relationship between the timing results may be determined at the second input/output processor. The first timing result may be sent over a dedicated communication channel between the input/output processors.

A plurality of timing criteria may be applied to a corresponding plurality of categories of data communications processed by the input/output processors to produce first and second pluralities of timing results. Relationships between corresponding ones of the first plurality of timing results and the second plurality of timing results may be determined.

An advantage offered by the invention is that failures in the communication path from the fault-tolerant system to the data network can be identified and in particular, a failure in the data paths coupling the input/output processors can be detected as distinct from a failure in a processor.

Other features and advantages of the invention will be apparent from the following description, including the drawings, and from the claims.

Brief Description of the Drawings

Fig. 1 is a block diagram of a fault-tolerant computing system with redundant computing elements and input/output processors.

Fig. 2 is a state diagram for an input/output processor.

Figs. 3-5 are flow charts of operations performed by an input/output processor.

Description

Referring to Fig. 1, a fault-tolerant system 10 includes dual-redundant compute elements (CEs) 12 and 14, dual-redundant input/output processors (IOPs) 24 and 26, and communication interconnection devices 36 and 38. CEs 12 and 14 carry out parallel operation sequences. Each CE communicates with both IOPs 24 and 26. CE 12 communicates over communication links 50 and 54, while CE 14 communicates over communication links 52 and 56. The communication links are routed through communication interconnection devices 36 and 38.

Each CE includes a central processing unit (CPU) 16 or 18 and an interface controller (IC) 20 or 22. The ICs provide an interface between the CPUs and the communication links. For example, an I/O request by CPU 16 is transmitted by IC 20 to IOPs 24 and 26 through communication links 50 and 54. With their interconnected communication structure, the IOPs expect to receive identical sequences of commands from each CE in normal operation.

Each IOP includes a CPU 28 or 30 and a network interface controller (NIC) 32 or 34. Network providers

29 and 31 are software drivers that execute on CPUs 28 and 30. NICs 32 and 34 allow the network providers to communicate over a local area network (LAN) 40 through network connections 62 and 64. A dedicated communication path 60 joining CPUs 28 and 30 allows network providers 29 and 31 to exchange messages without using LAN 40.

One network provider operates in a primary state while the other network provider operates in a standby state. Only the network provider operating in the primary state transmits data that originates in CE 12 or 14 to other devices on the LAN.

Network connections 62 and 64 connect to NICs 32 and 34 and terminate at ports of communication hubs 42 and 44 of the LAN 40. Hubs 42 and 44 are connected through a bridge 66 of LAN 40. Hubs 42 and 44 do not filter any communication while bridge 66 filters communication that is not destined to a device accessed through a particular port on the bridge. Hubs 42 and 44 are therefore on different segments of LAN 40. Other devices, such as a device 46, connected to the LAN 40 may communicate with fault-tolerant system 10.

Each of NICs 32 and 34 has a fixed, unique "physical" address and a programmable "logical" address that is configured to be the same for both NICs. The logical address is used for communication between the fault-tolerant system 10 and devices on LAN 40, such as device 46, or devices accessible from LAN 40. Each NIC is also programmed to receive group-addressed messages, such as messages sent to broadcast, multicast, or functional addresses. A group addressed message sent by a NIC specifies the NIC's unique physical address as the source of the message. As such, the recipient of a group addressed message can determine which NIC sent the message.

Network connections 62 and 64 may terminate on a common segment of LAN 40, or on different segments. In general, if connections 62 and 64 terminate on a common segment, then both the NICs may monitor all data traffic on that segment. Accordingly, data transmitted by one NIC may be received by the other NIC even if the data is not addressed to that NIC. By contrast, as shown in Fig. 1, network connections 62 and 64 may terminate on different segments at hubs 42 and 44. These hubs are connected by bridge 66 such that they are on different segments of LAN 40. Bridge 66 is configured to filter the data transmitted to a segment to avoid unnecessary use of communication capacity of that segment. A table of addresses of devices that are on a particular segment connected to bridge 66, or which communicate through that segment, is maintained by the bridge by monitoring data arriving from that segment. A message addressed to a specific device (i.e., a directed message instead of a group addressed message) that is not in the table for a segment is not retransmitted to that segment by bridge 66. On the other hand, group addressed messages are retransmitted on all segments of a LAN without filtering.

In operation, fault-tolerant system 10 determines whether network connections 62 and 64 are connected to a common segment of LAN 40 and therefore whether both NICs should expect to see identical network traffic. If the system determines that the NICs are on different segments, then the system determines that the NICs should expect to see only identical group addressed traffic.

In operation, both the primary network provider and the standby network provider monitor data communications to determine whether a fault has occurred. In the event that a fault would render the network connection of an IOP non-functional, appropriate action is taken. If the active network provider loses network

connectivity and the standby network provider is online, a switchover will occur to make the standby network provider the new primary network provider. If the standby network provider loses network connectivity, the standby network provider will enter an offline state until connectivity is reestablished.

The network providers detect faults by monitoring categories of data communication and maintaining the time since the last communication in each category occurred. When a network provider suspects that a communication failure may have occurred, it exchanges a status message with the other network provider over the communication path 60. The status message contains the times of last communication. Each network provider compares the times in a received status message to times maintained by the receiving network provider to identify, failures, if any, in the system. In making the comparisons, the network providers consider a tolerance within which the times should agree. This tolerance accounts for natural variability in transit times and for time needed to assemble and transmit the status messages.

To sense the status of network connections 62 and 64, as well as the status of LAN 40, network providers 29 and 31 periodically transmit through NICs 32 and 34 group addressed messages, known as noise packets, addressed to a group address monitored by both of the NICs. The source address of the group addressed message is set to the unique physical address of the transmitting NIC so that the receiving NICs can determine the source. When the network connections are on a common segment, the message is directly received by the NICs. When the source and receiving NICs are on separate segments, the group addressed packets are retransmitted from one segment to another in normal operation of LAN 40.

When fault-tolerant system 10 is initialized, network providers 29 and 31 go through a sequence of three startup states. In a first state, identified as the joined state, both IOPs have established communication with LAN 40, and have established communication over communication path 60 between processors 28 and 30. Next, in a synchronized state, communication redirection software executing on CPUs 16 and 18 is synchronized with network providers 29 and 31. Finally, in a fully initialized state, input/output requests executed on CPUs 16 and 18 can be sent successfully to network providers 29 and 31 for communication with LAN 40.

Referring to Fig. 2, when both network providers are fully initialized, both network providers enter an online/standby state 70. One network provider is subsequently taken from online/standby state to online/primary state 72. Whenever a network provider operating in one of the online states 70, 72 suspects that it may have lost network connectivity, the network provider periodically sends network status requests to the other network provider. If a loss of network connectivity is confirmed, the network provider goes to offline state 74. If the network provider operating in the online/primary state 72 detects the loss of network connectivity and goes to offline state 74, the network provider operating in the online/standby state 70 goes to online/primary state 72. Note that the automatic transition from online/primary state 72 to offline state 74 is only allowed when the other network provider is in online/standby state 70. While in offline state 74, a network provider periodically sends network status requests to the other network provider over link 60. If network connectivity is reestablished, then the network provider returns to online/standby state 70.

In addition to automatic transitions, operator controlled changes of state from online and offline states 70, 72, and 74 to a disabled state 76 may occur. When a network provider in state 76 is manually re-enabled, the network provider enters online/standby state 70 and then transitions immediately to offline state 74 if network connectivity is not confirmed. Finally, other detection mechanisms can determine that a network provider has failed, which results in the network provider entering a faulted state 78.

Referring to Fig. 3, when a network provider is in online or offline states 70, 72, or 74, the network provider repeatedly checks if a network status request should be sent to the other network provider over communication path 60. In particular the network provider determines whether to send a network status request upon expiration of an interval referred to as the NetworkStatusInterval. The default value for NetworkStatusInterval is 1000 milliseconds. Upon expiration of that interval, the network provider determines if one of three conditions is met. The first condition is true if no non-noise packets have been received in an interval referred to as ReceivePacketInterval (step 80). The default value of ReceivePacketInterval is 4000 milliseconds which corresponds to the typical maximum interval between packets received by system 10. The second condition is true if no noise packets have been received from the other network provider in the previous ReceivePacketInterval (step 82). The third condition is true whenever the network provider is in offline state 74 (step 84). If any of these conditions is met, a network status request is sent to the other network provider (step 86). Requests are transmitted without

consideration of whether a response has been received to a previous request.

Upon receipt of a network status request, a network provider typically constructs a response message containing the following communication statistics:

TimeLastNoiseReceived:	Time since the last noise packet was received from the other processor
TimeLastPacketReceived:	Time since the last non-noise packet addressed to the logical (system) address was received
TimeLastMulticastReceived:	Time since the last non-noise group addressed packet was received
TimeLastNoiseTransmitted:	The time since the last noise message was sent
TimeNetworkMonitored:	The time the network provider has been gathering statistics (i.e., the uptime of the system)
CountTransmitFailures:	Current transmit failure count

To protect against both network providers concurrently detecting failures and leaving no network provider in online/primary state 72, a network provider in an online/primary state 72 responds differently to a network status request than a network provider not in the online/primary state. If a local network provider in online/primary state 72 sends a network status request and receives a network status request from the remote network provider prior to receiving a response to its own request, the local network provider uses information in

the received request to satisfy its request, rather than waiting for a response. The local network provider does not reply to this received request. When a local network provider not in online/primary state 72 sends a network status request and receives a network status request from the remote network provider prior to receiving a response to its own request, the local network provider responds to the request and does not use the information in the received request to satisfy its pending request.

Referring to Fig. 4, upon receipt of a network status response (step 87), a local network provider determines whether the local or the remote network provider is not yet in the fully initialized state (step 88), if the remote network provider is in the faulted state 78 or the disabled state 76 (step 90), or if the local network provider is in the faulted state 78 or the disabled state 76 (i.e., not in online or offline states 70, 72 or 74) (step 92). If none of these conditions is true, the local network provider executes a procedure 94 to determine if the IOP has lost network connectivity. If any condition is true, the response is discarded (step 96) and no processing of the response is performed.

Referring to Fig. 5, the first step in procedure 94 is to determine whether both IOPs are receiving common traffic (step 100). The definition of common traffic depends on whether the local network provider determines that both IOPs are on a single segment or on different segments of LAN 40. Initially, both network providers assume that the IOPs are on different segments. In this case, common traffic corresponds to group addressed packets, other than noise packets, received by the IOPs. Both IOPs see common traffic if the value of TimeLastMulticastReceived in the network status response and the value computed at the local network provider are within a tolerance referred to as ReceiveTolerance. The

default value of ReceiveTolerance is 1000 milliseconds. If the network providers have determined that they are on the same segment, common traffic also includes packets directed to the logical address of system 10. Therefore, in addition to comparing TimeLastMulticastReceived, TimeLastPacketReceived is compared and both IOPs see common traffic if the times are within ReceiveTolerance.

If both IOPs do not see common traffic, the procedure determines if the local IOP is receiving common traffic (step 106). The local network provider determines that the local IOP does not see common traffic if the value of TimeLastMulticastReceived or TimeLastPacketReceived for the local IOP is greater than the received value by at least ReceiveTolerance (i.e., if the last group addressed message was received locally at least ReceiveTolerance earlier than at the remote IOP). If the local IOP does not see common traffic then the network provider has determined that there is a fault in the receive path from LAN 40 to the IOP. The network provider therefore makes a transition to offline state 74 (or remains in that state) (step 108).

If both IOPs receive common traffic (step 100), the network provider determines whether neither of the IOPs is receiving the other's noise packets (step 102). This occurs when the value of each network provider's TimeLastNoiseReceived is greater than the local value of TimeLastNoiseTransmitted by at least ReceiveTolerance, and the local value of TimeLastNoiseReceived is greater than the received value of TimeLastNoiseTransmitted by at least ReceiveTolerance. If neither IOP is receiving the other's noise packets (step 102), the local provider checks whether it is reporting a transmit failure (step 109). If not, the network provider assumes that the fault must be within LAN 40 because concurrent failure on connections 62 and 64 would be needed to account for the

status values. The simultaneous occurrence of both of these failure modes is assumed to be unlikely. If the network provide is reporting a transmit failure (step 109), then it goes to the offline state (step 108).

If at least one IOP is receiving the other's noise packets, the network provider determines whether locally transmitted noise packets are received by the other IOP (step 104). This occurs when the received value of TimeLastNoiseReceived does not exceed the local value of TimeLastNoiseTransmitted by at least ReceiveTolerance. If locally transmitted noise packets are received by the other IOP, then there is no fault with the local IOP. On the other hand, if the locally transmitted noise packets are not received by the other IOP, the network provider assumes that there is a fault in the transmit path from the local IOP to LAN 40, and the local network provider makes a transition to offline state 74 (step 108).

In all the tests, if a received time is greater than TimeNetworkMonitored, that time is considered invalid. This mechanism is used to prevent using inaccurate statistics. A further restriction on state changes is that a network provider in the offline state 74 must receive at least one packet while in that state before making a transition to online state 70. This restriction inhibits the state transition when there is absolutely no network traffic visible to either IOP.

Other embodiments are within the scope of the following claims. For example, the system described above uses dual redundant IOPs. Three or more IOPs can be used with a similar method of comparing the relative times of various categories of communications. When three or more IOPs are used, the responses from multiple IOPs can be used together to detect communication failures. In addition, relative times of other categories of system events than those described above

could be used for fault detection. Furthermore, the approach of using relative timing of communication events can be used to detect internal communication failures within the fault-tolerant system itself. Finally, the IOPs could be attached to different LANs if suitable forwarding of their noise packets were enabled.

What is claimed is:

1. A method for detecting a failure in a fault-tolerant computer system that includes a first input/output processor and a second input/output processor coupled to a data communication system, the method comprising the steps of:

applying a timing criterion to a category of data communications processed by the first input/output processor to produce a first timing result;

applying the timing criterion to the category of data communications processed by the second input/output processor to produce a second timing result;

determining a relationship between the first timing result and the second timing result; and

detecting whether a failure has occurred based on the determined relationship.

2. The method of claim 1 wherein the step of detecting whether the failure has occurred includes determining that a failure has occurred when a difference between the timing results exceeds a threshold value.

3. The method of claim 1 wherein the timing criterion is a time of last transmission or reception.

4. The method of claim 1 wherein the category of data communications includes messages originating from the first input/output processor.

5. The method of claim 4 wherein the category of data communications includes messages originating from the first input/output processor and directed to an address to which the second input/output processor is normally responsive.

6. The method of claim 4 wherein the category of data communications includes messages sent from the first input/output processor and directed through the data communication system to the second input/output processor.

7. The method of claim 1 wherein the category of data communications includes messages originating outside the computer system.

8. The method of claim 7 wherein the messages originate at a second computer coupled to the data communication system.

9. The method of claim 7 wherein the category of data communications includes messages originating outside the computer system and addressed to a group of systems of which the computer system is a member.

10. The method of claim 1 wherein the category of data communications includes messages originating from a third element of the computer system in data communication with the input/output processors.

11. The method of claim 1 further comprising the step of sending the first timing result from the first input/output processor to the second input/output processor wherein:

the step of applying the timing criterion to the category of data communications processed by the first input/output processor includes applying the timing criterion to the category of data communications processed by the first input/output processor at the first input/output processor;

the step of applying the timing criterion to the category of data communications processed by the second input/output processor includes applying the timing criterion to the category of data communications processed by the second input/output processor at the second input/output processor; and

the step of determining a relationship between the timing results includes determining a difference between the timing results at the second input/output processor.

12. The method of claim 11 wherein the first timing result is sent over a dedicated communication channel between the first input/output processor and the second input/output processor.

13. The method of claim 1 wherein

the step of applying a timing criterion to a category of data communications processed by the first input/output processor further includes applying a plurality of timing criteria to a corresponding plurality of categories of data communications processed by the first input/output processor to produce a first plurality of timing results;

the step of applying the timing criterion to the category of data communications processed by the second input/output processor further includes applying the plurality of timing criteria to the corresponding plurality of categories of data communications processed by the second input/output processor to produce a second plurality of timing results; and

the step of determining a relationship between the timing results further includes determining relationships between corresponding ones of the first plurality of timing results and the second plurality of timing results.

14. A fault-tolerant computer system coupled to a data communication system comprising:

a first input/output processor configured to process a category of data communications and to apply a timing criterion to the category of data communications to produce a first timing result; and

a second input/output processor configured to process the category of data communications and to apply a timing criterion to the category of data communications to produce a second timing result;

wherein the computer system is configured to determine a relationship between the timing results and to determine whether a failure has occurred based on the relationships.

15. The system of claim 14 wherein the timing criterion is a time of last transmission or reception.

16. The system of claim 14 wherein the category of data communications includes messages originating from the first input/output processor.

17. The system of claim 16 wherein the category of data communications includes messages sent from the first input/output processor and directed through the data communication system to the second input/output processor.

18. The system of claim 16 wherein the category of data communications includes messages originating from the first input/output processor and directed to an address to which the second input/output processor is normally responsive.

19. The system of claim 14 wherein the category of data communications includes messages originating outside the fault-tolerant system.

20. The system of claim 19 wherein the messages originate at a second computer coupled to the data communication system.

21. The system of claim 19 wherein the category of data communications includes messages originating outside the fault-tolerant system and addressed to a group of systems of which the fault-tolerant system is a member.

22. The system of claim 14 further comprising a third element of the computer system in data communication with the input/output processors, and wherein the category of data communications includes messages originating from said third element.

23. The system of claim 14 further comprising:

a dedicated communication channel coupling the first input/output processor and the second input/output processor, the communication channel being configured to send the first timing result from the first input/output processor to the second input/output processor;

wherein the second input/output processor is configured to determine a difference between the timing results and to detect whether the failure has occurred when the difference exceeds a threshold.

24. The system of claim 14 wherein:

the first input/output processor is further configured to apply a plurality of timing criteria to a corresponding plurality of categories of data

communications processed by the first input/output processor to produce a first plurality of timing results;

the second input/output processor is further configured to apply the plurality of timing criteria to the corresponding plurality of categories of data communications processed by the second input/output processor to produce a second plurality of timing results; and

the computer system is further configured to determine a relationships between corresponding ones of the first and second plurality of timing results and to determine whether the failure has occurred based on the determined relationships.

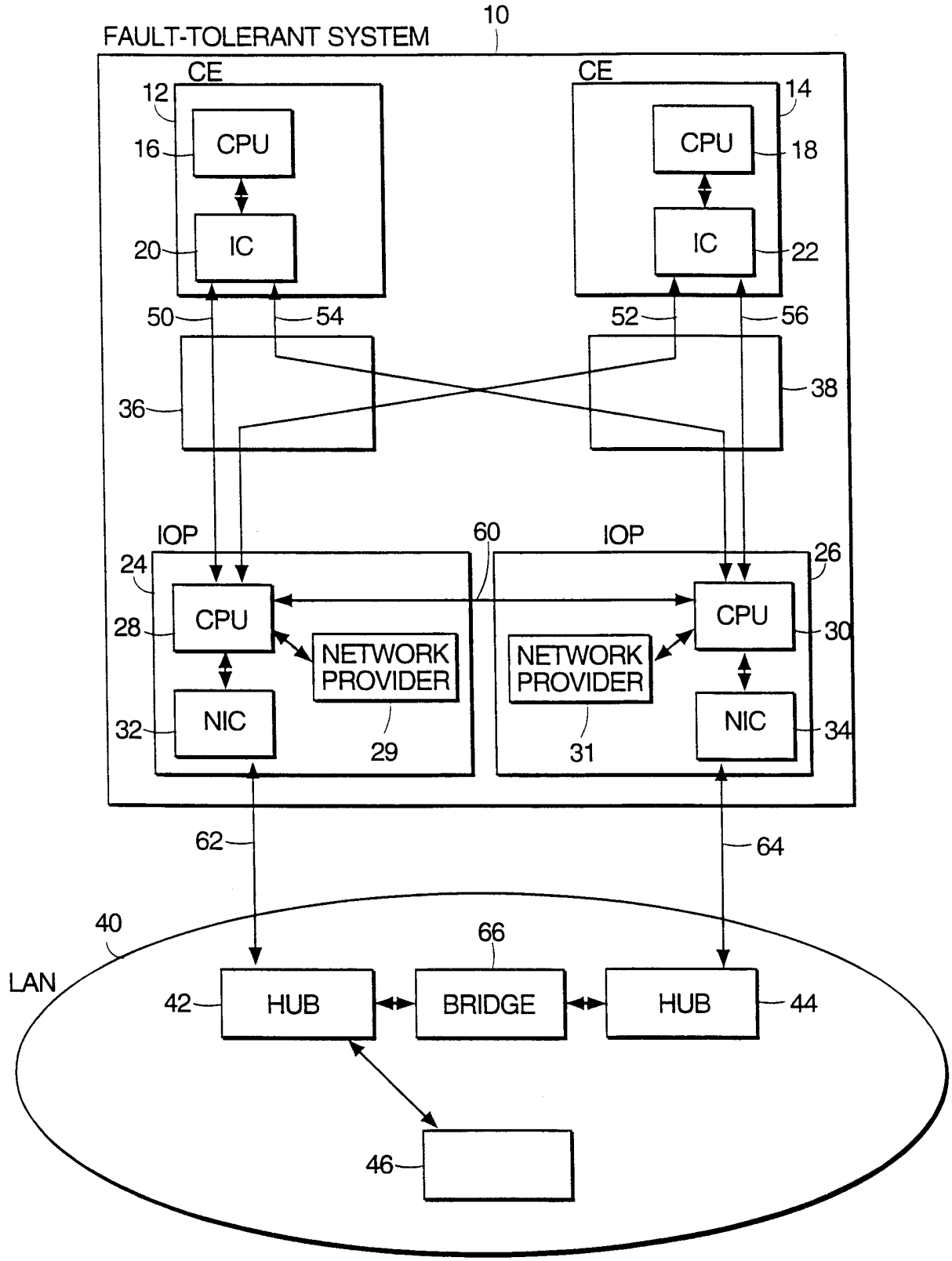

FAULT-TOLERANT SYSTEM
10
CE
CE
12
14
CPU
CPU
16
18
IC
IC
20
22
50
54
52
56
36
38
IOP
60
IOP
24
26
CPU
CPU
28
30
NETWORK PROVIDER
NETWORK PROVIDER
NIC
NIC
32
34
29
31
62
64
40
66
LAN
42
HUB
BRIDGE
HUB
44
46

FIG. 1

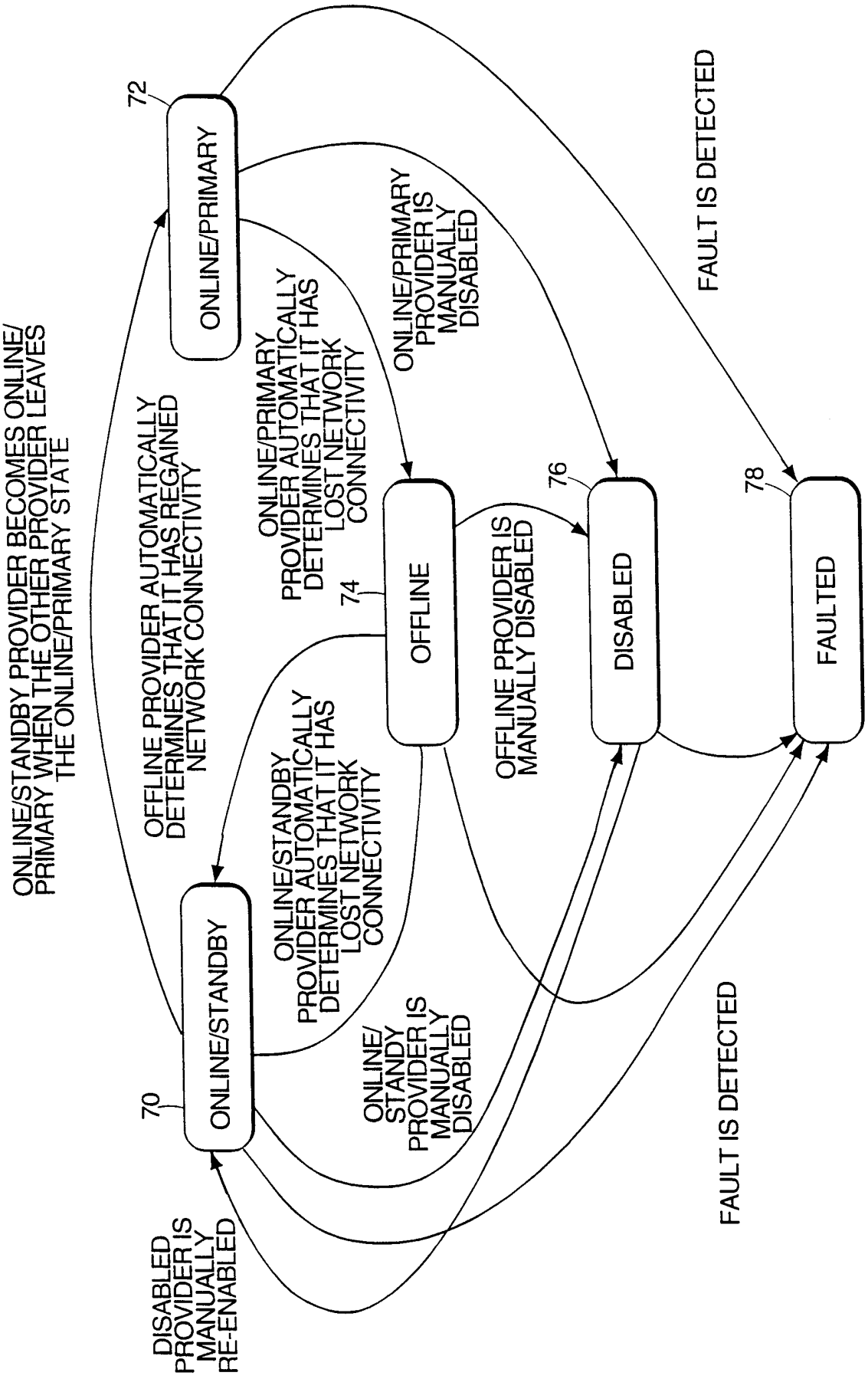

ONLINE/STANDBY PROVIDER BECOMES ONLINE/ PRIMARY WHEN THE OTHER PROVIDER LEAVES THE ONLINE/PRIMARY STATE
72
ONLINE/PRIMARY
FAULT IS DETECTED
ONLINE/PRIMARY PROVIDER IS MANUALLY DISABLED
ONLINE/PRIMARY PROVIDER AUTOMATICALLY DETERMINES THAT IT HAS LOST NETWORK CONNECTIVITY
OFFLINE PROVIDER AUTOMATICALLY DETERMINES THAT IT HAS REGAINED NETWORK CONNECTIVITY
74
OFFLINE
76
DISABLED
78
FAULTED
OFFLINE PROVIDER IS MANUALLY DISABLED
ONLINE/STANDBY PROVIDER AUTOMATICALLY DETERMINES THAT IT HAS LOST NETWORK CONNECTIVITY
ONLINE/ STANDY PROVIDER IS MANUALLY DISABLED
70
ONLINE/STANDBY
FAULT IS DETECTED
DISABLED PROVIDER IS MANUALLY RE-ENABLED

FIG. 2

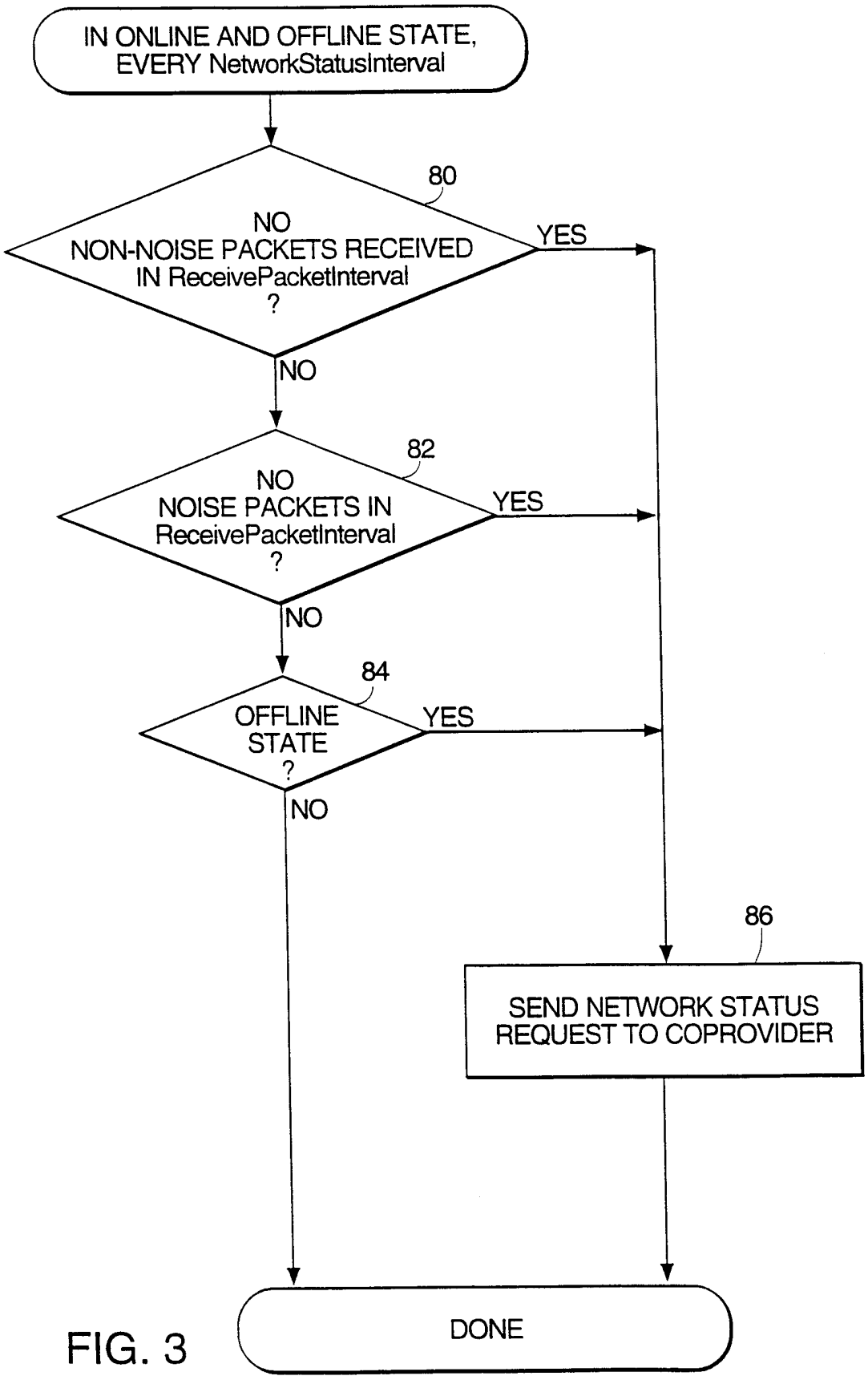

IN ONLINE AND OFFLINE STATE, EVERY NetworkStatusInterval
80
NO NON-NOISE PACKETS RECEIVED IN ReceivePacketInterval ?
YES
NO
82
NO NOISE PACKETS IN ReceivePacketInterval ?
YES
NO
84
OFFLINE STATE ?
YES
NO
86
SEND NETWORK STATUS REQUEST TO COPROVIDER
DONE

FIG. 3

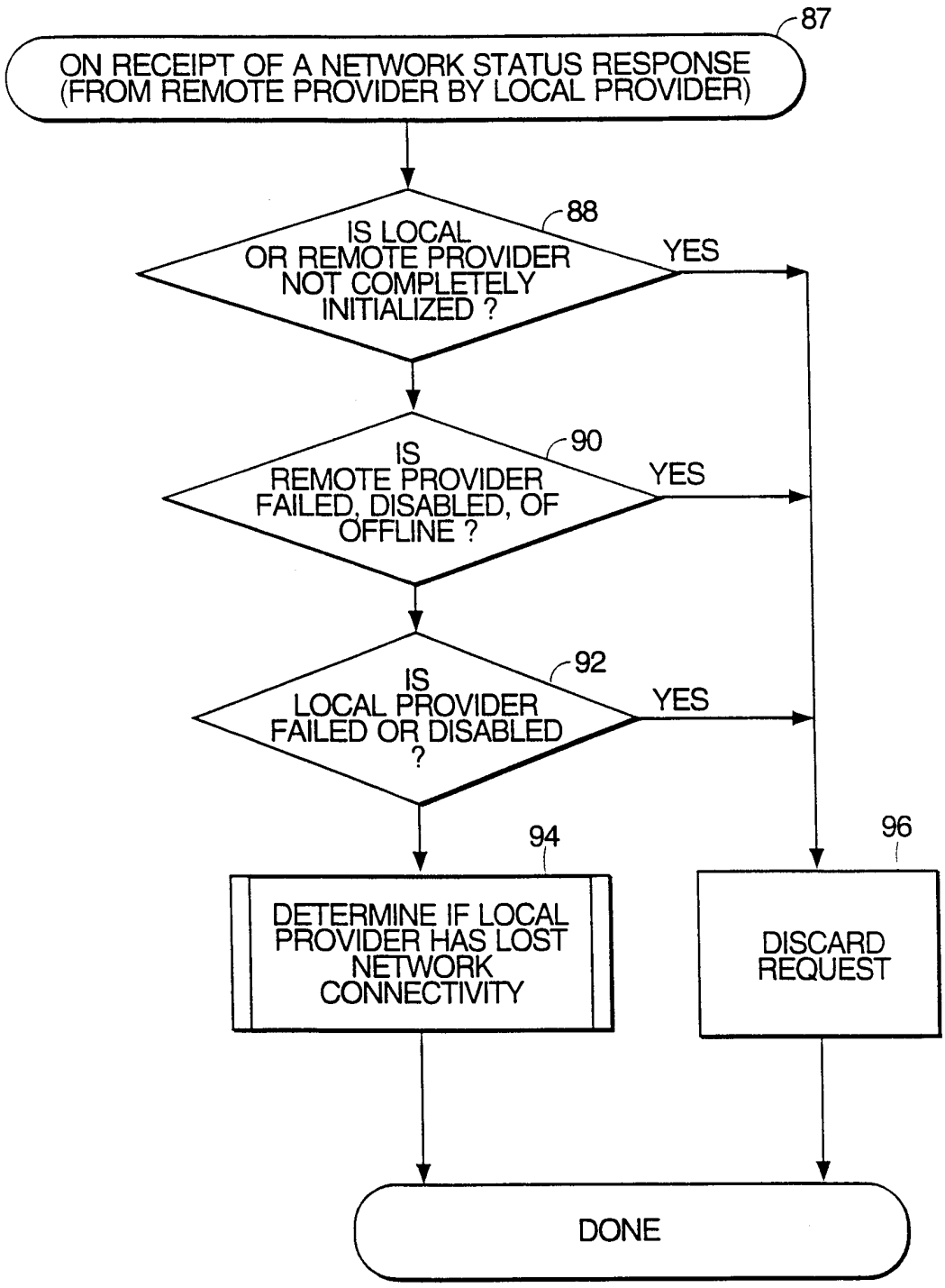

87
ON RECEIPT OF A NETWORK STATUS RESPONSE (FROM REMOTE PROVIDER BY LOCAL PROVIDER)
88
IS LOCAL OR REMOTE PROVIDER NOT COMPLETELY INITIALIZED ?
YES
90
IS REMOTE PROVIDER FAILED, DISABLED, OF OFFLINE ?
YES
92
IS LOCAL PROVIDER FAILED OR DISABLED ?
YES
94
DETERMINE IF LOCAL PROVIDER HAS LOST NETWORK CONNECTIVITY
96
DISCARD REQUEST
DONE

FIG. 4

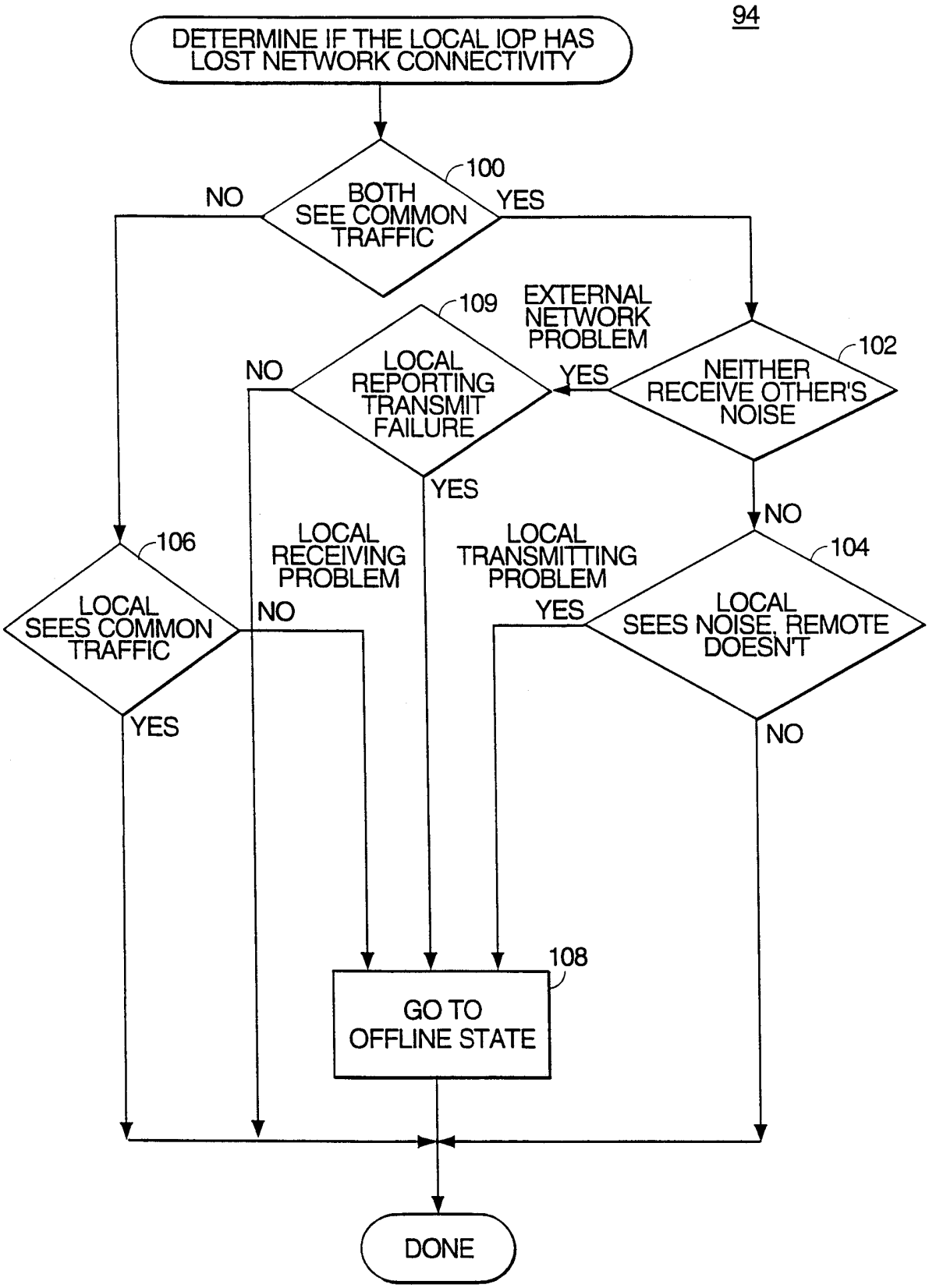

94
DETERMINE IF THE LOCAL IOP HAS LOST NETWORK CONNECTIVITY
100
BOTH SEE COMMON TRAFFIC
NO
YES
109
LOCAL REPORTING TRANSMIT FAILURE
EXTERNAL NETWORK PROBLEM
NO
YES
102
NEITHER RECEIVE OTHER'S NOISE
YES
NO
106
LOCAL SEES COMMON TRAFFIC
LOCAL RECEIVING PROBLEM
LOCAL TRANSMITTING PROBLEM
NO
YES
104
LOCAL SEES NOISE, REMOTE DOESN'T
YES
NO
108
GO TO OFFLINE STATE
DONE

FIG. 5

INTERNATIONAL SEARCH REPORT

I. ational Application No
PCT/US 98/14451

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F11/00 G06F11/20

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06F H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 0 649 092 A (TANDEM COMPUTERS INCORPORATED) 19 April 1995 see the whole document	1,14
A	EP 0 760 503 A (COMPAQ COMPUTER CORPORATION) 5 March 1997 see the whole document	1,14
A	US 4 610 013 A (LONG ET AL.) 2 September 1986 see abstract	1,14
A	US 4 710 926 A (BROWN ET AL.) 1 December 1987	
A	US 5 390 326 A (SHAH) 14 February 1995	

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
19 November 1998	30/11/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Absalom, R

1

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/US 98/14451

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 649092 A	19-04-1995	US 5448723 A	05-09-1995
		AU 685497 B	22-01-1998
		AU 7581694 A	04-05-1995
		CA 2117619 A	16-04-1995
		JP 2583023 B	19-02-1997
		JP 7235933 A	05-09-1995
EP 760503 A	05-03-1997	US 5696895 A	09-12-1997
		US 5781716 A	14-07-1998
US 4610013 A	02-09-1986	NONE	
US 4710926 A	01-12-1987	CA 1267226 A	27-03-1990
		DE 3688526 A	08-07-1993
		DE 3688526 T	09-12-1993
		EP 0230029 A	29-07-1987
		JP 2068920 C	10-07-1996
		JP 7104793 B	13-11-1995
		JP 62177634 A	04-08-1987
US 5390326 A	14-02-1995	NONE	